Credit Card Fraud Dataset Analysis - README

This document provides instructions and details on how to run the dataset analysis script. The script is designed to clean and analyse a dataset focused on credit card transactions. The key functions of the script include handling duplicates, calculating age based on date of birth (DOB), identifying missing data, and extracting unique job titles.

Prerequisites

1. Python Installation

- Python version 3.6 or higher must be installed in the system.

2. Python Libraries

The script requires the `pandas` library.

3. Dataset

- The dataset file should be a CSV file (`A1 Dataset - Dataset.csv`).
- Update the script to reflect the correct file path.

Running the Script

- **Download and Save the Dataset**

  Ensure that the dataset is available on your device and stored in the specified directory.

- **Script Overview**

  The script performs the following actions:

  - **Loads the dataset** into a Pandas DataFrame for analysis.
  - **Inspects the dataset** by displaying the first few rows, the shape of the dataset, and a summary of the features (data types, missing values, etc.).
  - **Handles duplicate rows** by checking for duplicates and providing the option to remove them.
  - **Extracts a subset of data** starting from the 924,850th row for further analysis.
  - **Resets the index** of the subset data to ensure clean indexing.
  - **Drops the last column** of the subset data as it is not needed for further analysis.

- **Calculates age** based on the dob (date of birth) column, using a reference date (2020-12-31). It ensures the dob is in the correct format and computes the age in years.
- **Identifies missing data** in the dataset by counting the number of missing values in each column.
- **Extracts unique job titles** from the `job` column for review.

- **How to Execute the Script**
  - Open a Python-compatible IDE or terminal.
  - Navigate to the directory where the script is saved.

- **Additional Steps**
  - If any issues arise (such as incorrect file paths or missing dependencies), ensure the dataset is correctly located, and the required libraries are installed.

- **Next Steps for Data Analysis**
  - After running the script, you can proceed to further analyze the cleaned dataset (`df_cut`).